ZARZĄDZANIE MEDIAMI
Tom 9(1)2021, s. 115–126
doi: 10.4467/23540214ZM.21.008.13056
www.ejournals.eu/ZM

Barbara Cyrek orcid.org/0000-0002-3270-6548

Wydział Zarządzania i Komunikacji Społecznej
Uniwersytet Jagielloński
cyrek.barbara@gmail.com

LIVE VIDEO STREAMING NA PLATFORMIE YOUTUBE – ZARZĄDZANIE TREŚCIAMI O CHARAKTERZE NIENAWISTNYM

Abstract

LIVE VIDEO STREAMING ON YOUTUBE – MANAGEMENT OF A HATEFUL CONTENT

The purpose of the article is to characterize current issues related to the occurrence of hate speech in live broadcasts on YouTube. Adding more tools to make the platform more attractive to users also gives new opportunities to spread hateful content. Platform policy must be constantly evolving, and subsequent regulations are widely commented on by both advertisers and users. For live broadcasts, real-time control requires automated solutions that are sometimes fallible.

Keywords: algorithmization, donations, Super Chat, live streams

Wprowadzenie

Choć mowa nienawiści stanowi ważne zagadnienie w dyskusji nad kondycją współczesnej kultury, nie ma zgody co do jednej wyczerpującej definicji tego zjawiska, która satysfakcjonowałaby zarówno praktyków, jak i przedstawicieli różnych dyscyplin naukowych. Warto zatem nakreślić, jaki sposób rozumienia mowy nienawiści został wybrany w danym opracowaniu. W niniejszej pracy przyjęto definicję pochodzącą z raportu Biura Studiów i Ekspertyz Kancelarii Sejmu RP, która za mowę nienawiści uznaje:

> (...) wypowiedzi ustne i pisemne oraz przedstawienia ikoniczne lżące, oskarżające, wyszydzające i poniżające grupy i jednostki z powodów po części od nich niezależnych – takich

> jak przynależność rasowa, etniczna i religijna, a także płeć, preferencje seksualne, kalectwo, czy przynależność do naturalnej grupy społecznej, jak mieszkańcy pewnego terytorium, reprezentanci określonego zawodu, mówiący określonym językiem. Jest to upubliczniona przemoc werbalna, wyraz nienawiści kolektywnej, adresowanej do zbiorowości naturalnych, wyznaczonych przez rasę, narodowość, płeć i wyznanie, do których nie przynależy się z racji swobodnie wybieranych przekonań (Łodziński, 2003, s. 5).

Z powyższego wynika, iż mową nienawiści jest nie tylko „słowo mówione", ale nacechowany pejoratywnie akt komunikacyjny, mający na celu deprecjonowanie wybranych, specyficznych grup. Nie bez znaczenia jest tutaj kwestia wspomnianego upublicznienia, gdyż właśnie ze względu na nie mowa nienawiści stanowi niebezpieczne zjawisko, które może implikować przeniesienie przemocy z warstwy werbalnej w sferę czynów.

Współczesne, zmediatyzowane społeczeństwo musi w szczególny sposób mierzyć się z omawianym zjawiskiem. Upowszechnienie technologii mobilnych poskutkowało wszechobecnością mediów. Użytkownicy mogą teraz za pomocą tabletów czy smartfonów nie tylko odbierać wiadomości bez ograniczeń czasu i miejsca, ale również sami być nadawcami komunikatów. Kultura partycypacji ma w tym kontekście dwojakie oblicze: owszem, ma potencjał aktywizacji jednostek i społeczności, ale aktywizacja ta nie jest ograniczona do działalności artystycznej czy prospołecznej i rozciąga się również na niepożądane społecznie zachowania – z mową nienawiści włącznie. Łatwość dostępu do nowych mediów[1] (w zakresie zarówno odbierania, jak i kreowania treści) implikuje nieustanne powiększanie się zasobów treści w internecie. Szacuje się, że internauci generują każdego dnia 2,5 kwintyliona bajtów danych (Petrov, 2019). Zjawisko nieustannego „stawania się" cyberprzestrzeni rodzi trudności w zakresie zwalczania mowy nienawiści. W tym kontekście warto zwrócić uwagę na fakt, że korporacje stojące za największymi platformami internetowymi definiują się jako „technologiczne", nie zaś „medialne". Kreując w ten sposób swoją tożsamość, organizacje te przypisują platformom rolę kanałów dystrybucji, a nie twórców treści, dzięki czemu częściowo zwalniają się z odpowiedzialności za występujące tam niepożądane zjawiska (Kreft, 2019, s. 67–77).

1. Metodologia

Artykuł ma na celu wykazanie i scharakteryzowanie najważniejszych aktualnych problemów wynikających z możliwości występowania mowy nienawiści w transmisjach emitowanych na żywo w serwisie YouTube. Przykład tej platformy pokazuje, że *live streaming* stał się wyzwaniem dla kontroli nadawanych treści – i to w skali globalnej.

[1] Rozumianych tutaj za Malwiną Popiołek jako media masowe nowej jakości zarówno w zakresie treści, formy, jak i dystrybucji przekazu. Zob. Popiołek, 2018, s. 18.

W celu scharakteryzowania tego problemu postawiono następujące pytania badawcze:

1. Jakie środki podejmuje Google w celu zapobiegania rozprzestrzenianiu się mowy nienawiści na platformie YouTube?
2. Jakie narzędzia do nadawania na żywo w serwisie YouTube mogą być wykorzystywane do publikowania nienawistnych treści?
3. Jakie kontrowersje towarzyszą wprowadzeniu zautomatyzowanych narzędzi cenzurowania treści?

Przedstawione w artykule wnioski są wynikiem badań przeprowadzonych w 2019 roku. W celu udzielenia odpowiedzi na pytania badawcze zastosowana została krytyczna analiza piśmiennictwa, w szczególności literatury medioznawczej, ale również doniesień prasowych z krajowych i zagranicznych portali. Badanie objęło również analizę polityki platformy YouTube. Okazało się to metodologicznym wyzwaniem, ponieważ wytyczne dla użytkowników i ogólne zasady panujące w witrynie aktualizowane są na bieżąco – brakuje archiwum zmian, w którym można byłoby prześledzić drobniejsze modyfikacje w zakresie wprowadzanej polityki (stąd też wiele źródeł nie ma podanych dat publikowania). Większe zmiany ogłaszane są natomiast na oficjalnym blogu (YouTube Official Blog) i dzięki temu daty ich wprowadzenia są łatwiejsze do odnalezienia. Całości analizy dopełniają studia przypadków youtuberów[2], którzy spotkali się z niesprawiedliwą lub niekonsekwentną cenzurą, a także tych, którzy doskonalą się w przechytrzaniu cenzurujących mechanizmów.

2. YouTube wobec mowy nienawiści

Główne przychody generowane przez platformę YouTube pochodzą ze sprzedaży powierzchni reklamowej i usług, co jest charakterystyczne dla serwisów społecznościowych (Kaznowski, 2008). Konieczność utrzymywania pozytywnych relacji z reklamodawcami jest jednym z powodów stosowania regulacji dotyczących charakteru publikowanych w serwisie treści. W tym zakresie polityka witryny YouTube nieustannie się rozwija – najważniejsze zmiany wprowadzone w ostatnich latach przedstawia tabela 1.

[2] Zapis za: Kosińska, 2014.

Tabela 1. Kluczowe zmiany polityki i produktów wprowadzone w serwisie YouTube od 2016 roku w celu zwalczania niestosownych treści

data	wprowadzone zmiany
2016	
17 listopada	Funkcjonalności pozwalające rodzicom blokować kanały na YouTube Kids
2017	
18 czerwca	Uczenie maszynowe i rozszerzenie oznaczania treści ekstremistycznych
31 lipca	Stworzenie z partnerami branżowymi *Global Internet Forum to Counter Terrorism*
1 sierpnia	Poszerzenie programu *Trusted Flagger* o organizacje pozarządowe
4 grudnia	Uczenie maszynowe i rozszerzenie oznaczania treści w celu wykrywania treści potencjalnie naruszających politykę związaną z mową nienawiści i bezpieczeństwem dzieci
2018	
styczeń	*Intelligence Desk* – zespół monitorujący wiadomości, media społecznościowe i zgłoszenia użytkowników w celu wykrywania nowych trendów związanych z niewłaściwymi treściami, zanim staną się one problemem na większą skalę
23 kwietnia	Opublikowanie pierwszego raportu *YouTube Community Guidelines Enforcement Report*; osiągnięcie większości wkładu w cel Google zakładający, że nad usuwaniem treści łamiących prawo będzie pracować 10 tysięcy osób
1 grudnia	Aktualizacja polityki zabraniającej małoletnim samotnego nadawania na żywo bez opieki dorosłych
2019	
15 stycznia	Aktualizacja wytycznych wyjaśniająca, kiedy treści zawierające żarty lub wyzwania łamią *Wytyczne dla społeczności*
19 lutego	Wprowadzenie nowego systemu *Community Guidelines strikes*, w tym ostrzeżeń i większej transparentności
28 lutego	Wyłączenie komentarzy dla materiałów wideo przedstawiających małoletnich w ryzykownych sytuacjach
5 czerwca	Aktualizacja polityki zakazującej treści poświęconych supremacji oraz o charakterze nienawistnym
1 lipca	Domyślna funkcja automatycznie wstrzymująca potencjalnie niewłaściwe komentarze do oceny twórców
21 sierpnia	Aktualizacja polityki zakazująca stosowania tytułów lub opisów kierowanych do rodzin w przypadku treści dla dojrzałych widzów

Źródło: YouTube Official Blog, The Four Rs of Responsibility, Part 1: Removing harmful content (2019).

Prócz obowiązujących regulaminów i warunków korzystania YouTube wprowadził również *Wytyczne dla społeczności* – zbiór dwunastu zasad, napisanych przystępnym językiem, które mają pomóc użytkownikom w zrozumieniu obowiązujących tam norm. Jeszcze przed charakterystyką poszczególnych zasad YouTube poucza: „Potraktuj je poważnie i weź sobie do serca. Nie szukaj w nich luk ani nie próbuj obchodzić ich za pomocą kruczków prawnych – po prostu postaraj się zrozumieć te zasady i postępować zgodnie z ich duchem" (YouTube, b.d., a). Jedna z wytycznych dotyczy „treści szerzących nienawiść" i brzmi następująco:

> Nasze usługi zapewniają swobodę wypowiedzi. Nie dotyczy to jednak materiałów, które zachęcają do przemocy wobec osób lub grup z powodu rasy, pochodzenia etnicznego, religii, niepełnosprawności, płci, wieku, statusu weterana bądź orientacji seksualnej czy tożsamości płciowej, a także tych, które szerzą nienawiść na tle wspomnianych cech. Nie zawsze jest to oczywiste, jednak jeśli głównym celem materiału jest atakowanie chronionych grup, jest on niedopuszczalny (YouTube, b.d., a).

Jest to skondensowany opis polityki serwisu YouTube wobec mowy nienawiści. Powyższe reguły są szerzej opisane i uzupełnione w *Zasadach dotyczących szerzenia nienawiści*, gdzie za mowę nienawiści uważa się również promowanie przemocy lub nienawiści wobec grup lub osób ze względów takich jak kasta, narodowość, status imigranta, bycie ofiarą brutalnego aktu przemocy lub osobą z nią spokrewnioną (YouTube, b.d., b). YouTube zaznacza przy tym, że zezwala na publikację takich treści, o ile mają one edukacyjny, dokumentalny, naukowy lub artystyczny charakter (YouTube, b.d., b). Niezwykle istotne, również w kontekście rozważań w niniejszym artykule, jest to, że wprowadzone przez serwis YouTube zasady dotyczą nie tylko materiałów wideo, ale treści w ogóle.

Mowa nienawiści w serwisie YouTube przyjmuje różną postać. Może ona występować w takich elementach jak: materiały wideo (filmy i relacje), posty na karcie „Społeczności" (zdjęcia i wpisy), zdjęcia (miniatura lub tło kanału), komentarze, transmisje na żywo, czat, donacje.

Sposoby zwalczania mowy nienawiści są zróżnicowane w zależności od jej źródła (twórca, widz) oraz tego, kto takiego przeciwdziałania się podejmuje (użytkownicy, algorytmy, zespół serwisu YouTube). Na przykład w przypadku komentarzy i wpisów na czacie podstawową funkcją jest usuwanie, które może być przeprowadzone na podstawie: a) oceny właściciela kanału lub moderatora czatu; b) zgłoszeń innych użytkowników; c) aktywnego blokowania słów. Ostatnia funkcja musi być nieustannie rozwijana, gdyż użytkownicy wymyślają coraz to nowsze sposoby na oszukanie „cenzurującego" algorytmu, np. poprzez stosowanie cyfr w miejsce konkretnych liter. Twórca ma również możliwość nałożenia na wybranych przez siebie użytkowników tak zwanego *shadow bana*, który polega na zablokowaniu danej osoby w sposób dla niej niewidoczny – widzi ona własne wpisy, które pozostają ukryte dla innych. Bardziej restrykcyjną formą przeciwdziałania szerzeniu niepożądanych treści jest zablokowanie możliwości komentowania filmu. Czasami twórcy

decydują się na taki ruch, kiedy dyskusja pod ich filmem obiera niechciany kierunek. Pozwala to ukryć również dodane wcześniej komentarze.

W przypadku, gdy sam materiał wideo łamie obowiązujące w serwisie zasady, YouTube może podjąć mniej lub bardziej restrykcyjne działania. W zależności od różnych czynników, takich jak np. historia przewinień danego użytkownika, serwis stosuje system ostrzeżeń – *warning*, *1st strike*, *2nd strike*, *3rd strike* (YouTube, b.d., c), które wiążą się z blokowaniem niektórych funkcji na określony regulaminem czas, a jeśli te nie przynoszą rezultatów, następuje trwałe usunięcie kanału. YouTube może także zadecydować o blokowaniu możliwości komentowania, usunięciu danego filmu, nałożeniu na film ograniczeń wiekowych i ostrzeżeń o charakterze przedstawianych tam treści (np. o drastycznych scenach). Specyficzną sankcją nakładaną przez platformę jest demonetyzacja, czyli wyłączenie możliwości zarabiania na danej produkcji. W praktyce oznacza to, że popularność materiału nie przekłada się na zarobki twórcy, który pozbawiony zostaje przychodu z opublikowanego filmu.

Działania podejmowane przez serwis YouTube w dużej mierze opierają się na uczeniu maszynowym i pracy algorytmów, które jednak nie są nieomylne. Im bardziej YouTube zaostrza politykę przeciwdziałania mowie nienawiści, tym więcej materiałów zostaje zablokowanych przez algorytmy – niekoniecznie słusznie. Ofiarą algorytmicznej cenzury padł m.in. nauczyciel historii Scott Allsop, pokazujący w materiałach edukacyjnych wypowiedzi Adolfa Hitlera i Josepha Goebbelsa. Filmy zostały usunięte, a z uzasadnienia takiej decyzji nauczyciel dowiedział się, że promowały one nienawiść lub przemoc wobec członków chronionej grupy. Materiały zostały ostatecznie przywrócone przez zespół YouTube'a[3] po zgłoszeniu zażalenia (Tyan, 2019).

Jak się okazuje, kontrola mowy nienawiści w serwisie bywa nie tylko omylna, ale również niekonsekwentna. Polski wideobloger Maciej Dąbrowski publikuje w serwisie YouTube serię filmów o tematyce kulturalnej. W każdej produkcji prezentowany jest jeden film, serial, gra, książka lub komiks oraz album muzyczny (ewentualnie wykonawca), które łączy przewodni temat danego materiału. Szesnasty odcinek cyklu, o tytule *NA CZARNO*, został przez algorytmy YouTube'a najpierw mocno ograniczony przez wyłączenie funkcji takich jak: „dodawanie komentarzy, udostępnianie, polecane filmy" (Dąbrowski, 2018a), następnie przywrócony – ale bez monetyzacji (Dąbrowski, 2018b), a na koniec umieszczony na liście polecanych i najpopularniejszych filmów – „Na czasie" (Dąbrowski, 2018c). Przykłady takie można oczywiście mnożyć i być może warto zgromadzić je w odrębnej publikacji i poddać analizie. W rozważaniach nad mową nienawiści w mediach społecznościowych należy jednak mieć na uwadze, że ogromny przyrost danych wiąże się z koniecznością zautomatyzowanej kontroli, stąd też algorytmy, choć omylne, są nadal niezbędne.

[3] Deklinacja za: Grzenia, 2009.

Ważną rolę w procesie walki z mową nienawiści odgrywają również sami użytkownicy. Mogą oni zgłaszać treści w ich ocenie szkodliwe, pojawiające się w każdej z wymienionych wyżej form wypowiedzi, bez względu na to, czy jest nią tekst, obraz czy materiał audiowizualny. Taka funkcjonalność pozwala społeczności serwisu na swoistą samoregulację, wzmacnianą działaniem algorytmów. Interesujący w tym względzie jest jednak fakt, że zgłoszeń treści potencjalnie niewłaściwych mogą dokonywać wyłącznie zalogowani użytkownicy. Ci zaś w 2013 roku stanowili zaledwie około 17% wszystkich widzów serwisu (Lee, 2013). Choć oglądanie materiałów na platformie YouTube jest względnie darmowe – widz płaci swoim czasem poświęconym na oglądanie reklam – to już dostęp do narzędzi służących przeciwdziałaniu mowie nienawiści należy „wykupić" za swoje dane, podawane podczas rejestracji w serwisie.

Wytyczne dla społeczności informują również, że zachowanie użytkowników poza YouTube'em może skutkować nałożeniem sankcji w serwisie: „Jeśli twórca YouTube swoim zachowaniem na platformie i poza nią szkodzi użytkownikom, społeczności, ekosystemowi lub naszym pracownikom, możemy zareagować, podejmując działania określone na podstawie różnych czynników" (YouTube, b.d., a).

3. Mowa nienawiści na żywo

W kwietniu 2011 roku YouTube ogłosił wprowadzenie powszechnie dostępnej funkcji nadawania transmisji na żywo (YouTube Official Blog, 2011). Informacja ta spotkała się z entuzjastycznym odbiorem użytkowników. YouTube w ich odbiorze jest – w przeciwieństwie do telewizji – medium bardziej autentycznym, którego twórcy „nie pudrują faktów" (Hatalska, 2016), a „nażywość" dodatkowo zwiększa ten potencjał, gdyż możliwości kontroli (w tym także cenzury) są w tym przypadku znacznie ograniczone.

Transmisje prowadzone na żywo stały się nowym wyzwaniem w walce z mową nienawiści. W przypadku, gdy nienawistne komunikaty są przedmiotem samej transmisji, YouTube podejmuje standardowe kroki, takie jak ostrzeżenia czy usunięcie kanału, a także usunięcie zapisu z transmisji (który może być archiwizowany w serwisie). O potencjalnie niedozwolonej treści takiego przekazu informuje zgłaszająca społeczność lub algorytmy pracujące nad wykrywaniem niedozwolonych treści. Sytuacje takie, choć trudniejsze do wykrycia niż w przypadku tradycyjnych materiałów wideo, nie implikują dylematów w związku z odpowiedzialnością za stosowanie mowy nienawiści – twórca, konkretny youtuber, zostaje ukarany za łamanie warunków korzystania z platformy. Osoby ukarane przez YouTube'a usunięciem kanału znalazły jednak sposób na to, by nadal być obecne w serwisie – pojawiają się jako goście w transmisjach nadawanych na żywo przez innych użytkowników (CNET News staff, 2018). Proceder ten nie jest łatwy do ukrócenia,

a platforma zmuszona jest tutaj balansować na cienkiej granicy pomiędzy wolnością słowa a mową nienawiści.

Zjawiskiem jeszcze bardziej kontrowersyjnym jest występowanie treści o charakterze nienawistnym na czacie towarzyszącym transmisji na żywo. Czat może być dostępny dla wszystkich użytkowników lub ograniczony wyłącznie do grona osób wspierających kanał[4]. YouTube oferuje dwa widoki czatu: topczat, filtrujący wiadomości, takie jak potencjalny spam, oraz czat na żywo, wyświetlający na bieżąco wszystkie wiadomości. Jak informuje YouTube: „Komentarze widzów mogą mieć nie więcej niż 200 znaków. Można dodać maksymalnie 3 wypowiedzi co 30 sekund. Właściciel kanału nie podlega temu limitowi częstotliwości. Widzowie nie mogą dodawać znaków specjalnych, adresów URL ani tagów HTML" (YouTube, b.d., d). Istnieje możliwość ponownego odtworzenia czatu w przypadku wyświetlania zarchiwizowanych transmisji.

Odpowiedzialność za treści nienawistne publikowane na czacie jest kwestią dyskusyjną. Jeśli autorem wiadomości nie jest nadawca transmisji na żywo, czy można obarczyć go konsekwencjami za pojawiającą się na czacie mowę nienawiści? Wielu twórców, czy to w trosce o dobro społeczności, czy w obawie przed sankcjami, decyduje się na aktywne blokowanie wybranych słów, korzystając ze wsparcia algorytmów cenzurujących. Inną, opcjonalną formą pomocy ze strony algorytmów jest przesyłanie do sprawdzenia wiadomości potencjalnie nieodpowiednich. Istnieje również możliwość nadania wybranym użytkownikom statusu moderatorów, dzięki czemu mogą oni, na podstawie własnej oceny, usuwać wpisy łamiące politykę obowiązującą w serwisie oraz nakładać ograniczenia na użytkowników publikujących takie wpisy (zgłaszanie do YouTube'a, zawieszanie, ukrywanie). YouTube rekomenduje, by podczas wydarzeń o dużej popularności korzystać z pomocy co najmniej jednego moderatora czatu (YouTube, b.d., d).

Problematyka występowania mowy nienawiści w wiadomościach publikowanych na czacie skomplikowała się jeszcze bardziej na początku 2017 roku wraz z wprowadzeniem narzędzia, jakim jest superczat (ang. *Super Chat*). Pozwala on na przekazanie darowizny pieniężnej (donacji) z poziomu czatu, wraz z widoczną publicznie wiadomością tekstową. System „nagradza" donatorów, wyświetlając ich wiadomości na górze okna czatu. Wiadomości te mogą być wyświetlane na ekranie transmisji na żywo oraz odczytywane przez symulator głosu.

Narzędzie to, nazywane „słoikiem na napiwki" (ang. *tip jar*) (Solsman, 2017) dla youtuberów, nie jest wyłącznie wyrazem dobrej woli platformy, której funkcjonowanie opiera się m.in. na dzieleniu zysków z twórcami. Google pobiera 45% zysków z reklam wyświetlanych z materiałami wideo, youtuberom pozostawiając pozostałe 55% (Rosenberg, 2018). Jak zostało już nakreślone w niniejszym artykule, relacje firmy z reklamodawcami są kluczowe dla generowania zysków przez

[4] Status wspierającego uzyskuje się przez dokonanie comiesięcznych płatności. Wysokość opłaty może być różna w zależności od takich czynników jak np. kraj czy region.

serwis YouTube. Przedsiębiorstwa wykupujące reklamy na platformie wyrażają jednak obawy co do tego, jakie treści towarzyszyć będą ich materiałom promocyjnym. W związku z tym YouTube postanowił zarabiać również na twórcach, których działalność niekoniecznie uznawana jest za przyjazną reklamodawcom (ang. *advertiser-friendly*). Sposobem na to jest właśnie superczat, który dzieli donację pomiędzy beneficjenta oraz samą platformę. YouTube zatrzymuje tutaj 30% kwoty (Baker, 2018).

Donacje okazały się zupełnie nowym wyzwaniem w walce z mową nienawiści. Niektórzy użytkownicy przyjęli, że uiszczenie opłaty pozwala im na bezkarne szerzenie nienawistnych treści, a mechanizm algorytmicznego „nagradzania" za przekazaną kwotę tylko pogłębił ten problem. Zjawisko szerzenia ideologii rasistowskiej i nazistowskiej poprzez donacje zwróciło uwagę dziennikarzy. Zdaniem Vlada Savova (2017) z portalu The Verge:

> Możesz wybrać najbardziej niewinne tematy, takie jak – powiedzmy – kocięta drzemiące w wiklinowym koszu w słoneczny dzień, i nie zajmie ci dużo czasu znalezienie niezwykle upartego trolla spamującego czat nazistowskimi insygniami, rasistowskim lżeniem lub czymkolwiek innym zaprojektowanym, by szokować i obrażać jak najwięcej ludzi. Tak, czaty na YouTubie wymagają rozwiązania, ale pomysł Google na superczat jest najgorszym możliwym rozwiązaniem tego problemu[5].

O problemie nienawiści w donacjach skierowanej przeciwko mniejszości żydowskiej pisano m.in. również w „The Hill" (Bowden, 2018), „The Wall Street Journal" (Koh, 2018) czy „BuzzFeed News" (Daro, Silverman, 2018). Mimo to zjawisko to nie jest szeroko komentowane w środowisku naukowym, zwłaszcza na polskim gruncie.

Donatorzy, jak się okazało, nie tylko wykorzystują superczat do publikowania własnych poglądów i przekonań, ale również nakłaniają youtuberów do podejmowania różnych działań podczas transmisji. Zdaniem Joan Donovan, „kiedy połączysz transmisję i natychmiastową monetyzację, otrzymujesz wzór zachowań coraz bardziej zmierzający w kierunku skrajności" (Daro, Silverman, 2018).

W raporcie *Alternative Influence: Broadcasting the Reactionary Right on YouTube* Rebecca Lewis (2018) zwraca uwagę na szerszy kontekst zjawiska. Superczat dostarcza także zautomatyzowanych narzędzi do wpływania na kształt samej transmisji emitowanej na żywo. Automatyczne odczytywanie treści wiadomości sprawia, że donator „zabiera głos". Ofiarą takiego zjawiska padł m.in. kierowca Ubera, który 15 marca 2018 roku wiózł trójkę nadających na żywo youtuberów i zmuszony był wysłuchiwać obraźliwych donacji, m.in. dotyczących tematu pedofilii (Lewis, 2018, s. 41–42). Lewis uważa, że system natychmiastowej monetyzacji w transmisjach na żywo zachęca zarówno poszczególnych influencerów[6], jak i serwis YouTube (jako

[5] Cytaty z dzieł obcojęzycznych w tłumaczeniu autorki.

[6] Zapis za: Burzyński, 2016.

podmiot korporacyjny) do czerpania pieniędzy z ekstremistycznych treści publikowanych przez publiczność (Lewis, 2018, s. 42).

Jednak YouTube postanowił odciąć się od zarabiania na mowie nienawiści publikowanej za pomocą superczatu. W artykule opublikowanym w „The Guardian" we wrześniu 2018 roku (dokładnie tego samego dnia, kiedy opublikowano raport Lewis, który zresztą jest tam szeroko komentowany) przeczytać można, że:

> (...) firma zaostrzyła zasady, dzięki którym kanały mają dostęp do funkcji monetyzacji, i wdrożyła technologię uczenia maszynowego, aby identyfikować mowę nienawiści w funkcjach komentowania. Jeśli użytkownik zostanie przyłapany na wysyłaniu obraźliwych superczatów, otrzymane przychody zostaną przekazane na cele charytatywne, a nie podzielone między serwis YouTube a twórcę kanału (Solon, 2018).

Informacja ta poruszyła środowisko youtuberów, ponieważ technologia (a cytat ten wskazuje, że to ona ma być odpowiedzialna za cenzurowanie donacji) bywa omylna i „nie rozumie" sarkazmu czy humoru. Sprawa jest o tyle poważna, że nie dotyczy już tylko wirtualnych wartości (takich jak polubienia czy wyświetlenia), ale realnych pieniędzy. Youtuberzy zwrócili uwagę, że rozwiązanie takie nadal jest korzystniejsze dla samego YouTube'a niż dla twórców – firma może liczyć na ulgi podatkowe z tytułu dokonania darowizny, a więc i tak zyskuje na sytuacji (KEEM, 2018). Choć pierwotnie szeroko komentowana w mediach, sprawa po jakimś czasie ucichła. Aktualnie, w drugiej połowie 2019 roku, w centrum pomocy serwisu YouTube widnieje informacja:

> Podobnie jak wszystko w YouTubie, wysyłane superczaty muszą być zgodne z naszymi *Wytycznymi dla społeczności*. Jeśli superczat zostanie poddany moderacji i usunięty z powodu naruszenia naszych zasad, YouTube przekaże naszą część dochodów na cele charytatywne (YouTube, b.d., e).

Podsumowanie

Mowa nienawiści, rozumiana jako upubliczniony akt komunikacyjny propagujący bądź uzasadniający przemoc wobec osób lub grup, może przyjąć różnoraką postać w mediach społecznościowych – od tekstowej po audiowizualną. Serwis YouTube, szczególnie ceniący relacje z reklamodawcami, opracował zestaw warunków korzystania z oferowanych usług, pośród których znajdują się szczegółowe zasady dotyczące przeciwdziałania mowie nienawiści. Wraz z rozwojem witryny pojawiały się coraz to nowsze wyzwania na tym polu, a polityka platformy ulegała rozszerzeniu.

Szczególnie trudną do kontrolowania sferą w obrębie funkcjonowania serwisu są transmisje emitowane na żywo. Stanowią one niełatwy do ocenzurowania materiał, a walka z mową nienawiści z konieczności wymaga tu automatyzacji. Algorytmy, mimo omylności, a czasem również niekonsekwencji, stały się nieodzownym stróżem poprawności.

Pojawienie się w 2017 roku superczatu dodatkowo skomplikowało walkę z nienawistnymi treściami, zaś rozwiązania wprowadzone przez YouTube'a wzbudziły kontrowersje w środowisku twórców. Zasady przechwytywania donacji oraz ich dalszego losu nadal wydają się nieprzejrzyste. Czy można odwołać się od decyzji „cenzora"? Czy w przyszłości uda się opracować nieomylny algorytm, którego działanie będzie satysfakcjonować zarówno zespół platformy, jak i społeczność użytkowników? Które organizacje zostaną beneficjentami przejętych przez YouTube'a donacji? Czy istnieje możliwość zastąpienia algorytmicznej cenzury przez moderację samego twórcy? A jeśli tak – jakie wówczas ponosiłby konsekwencje niedostatecznego przeciwdziałania mowie nienawiści? Choć na niektóre z tych pytań nie można dzisiaj dać jednoznacznej odpowiedzi, warto temat ten poruszać w naukowej dyskusji, mając przy tym na uwadze, jak cienka jest granica pomiędzy mową nienawiści a wolnością słowa.

Bibliografia

Baker A. (2018). *Make Money with YouTube Super Chat*, http://contentcareer.com/blog/make-money-with-youtube-super-chat/ (dostęp: 24.09.2019).

Bowden J. (2018). *YouTube's paid comment feature being used to promote hate speech: Report*, https://thehill.com/policy/technology/388115-youtubes-paid-comment-feature-being-used-to-promote-hate-speech-report (dostęp: 24.09.2019).

Burzyński J. (2016). *Influencer*. Obserwatorium Językowe Uniwersytetu Warszawskiego, http://nowewyrazy.uw.edu.pl/haslo/influencer.html (dostęp: 27.09.2019).

CNET News Staff (2018). *YouTube's Super Chat hijacked for hate speech*, https://www.cnet.com/news/youtubes-super-chat-hijacked-for-hate-speech/ (dostęp: 26.09.2019).

Daro I.N., Silverman C. (2018). *How YouTube's "Super Chat" System Is Pushing Video Creators Toward More Extreme Content*, https://www.buzzfeednews.com/article/ishmaeldaro/youtube-comments-hate-speech-racist-white-nationalists-super?mod=article_inline (dostęp: 20.09.2019).

Dąbrowski M. (2018a). *No to YT zaszalał*, https://www.facebook.com/zdupy/photos/a.393173784147816/1215013588630494/?type=3&theater (dostęp: 20.09.2019).

Dąbrowski M. (2018b). *VICTORY HAS DEFEATED YOU!*, https://www.facebook.com/zdupy/photos/a.393173784147816/1216177608514092/?type=3&theater (dostęp: 20.09.2019).

Dąbrowski M. (2018c). *Ahahaha i jeszcze po odblokowaniu załapało się do Na Czasie*, https://www.facebook.com/zdupy/photos/a.393173784147816/1216288968502956/?type=3&theater (dostęp: 20.09.2019).

Grzenia J. (2009). *YouTube*, https://sjp.pwn.pl/poradnia/haslo/YouTube;10705.html (dostęp: 28.09.2019).

Hatalska N. (2016). *Rola blogerów i youtuberów we współczesnym świecie*, https://blogforumgdansk.pl/downloads/Raport_Rola_blogerow_youtuberow_we_wspolczesnym_swiecie_BFGdansk.pdf (dostęp: 24.09.2019).

Kaznowski D. (2008). *Nowy marketing*. Warszawa: VFP Communications Sp. z o.o.

KEEM (2018). *If youtube deems your superchat to be offensive…*, https://twitter.com/KEEMSTAR/status/1053586469829443584 (dostęp: 27.09.2019).

Koh Y. (2018). *Hate Speech on Live 'Super Chats' Tests YouTube*, https://www.wsj.com/articles/hate-speech-on-live-super-chats-tests-youtube-1541205849 (dostęp: 20.09.2019).
Kosińska K. (2014). *Youtuber*. Obserwatorium Językowe Uniwersytetu Warszawskiego, http://nowewyrazy.uw.edu.pl/haslo/youtuber.html (dostęp: 7.09.2019).
Kreft J. (2019). *Władza algorytmów. U źródeł potęgi Google i Facebooka*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.
Lee Y. (2013). *What percentage of people viewing YouTube videos on YouTube.com are logged into the service?*, https://www.quora.com/What-percentage-of-people-viewing-YouTube-videos-on-YouTube-com-are-logged-into-the-service (dostęp: 22.09.2019).
Lewis R. (2018). *Alternative Influence: Broadcasting the Reactionary Right on YouTube*. Data & Society Research Institute, https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf (dostęp: 27.09.2019).
Łodziński S. (2003). *Problemy dyskryminacji osób należących do mniejszości narodowych i etnicznych w Polsce*. Wydział Analiz Ekonomicznych i Społecznych Kancelarii Sejmu, Biuro Studiów i Ekspertyz. Raport nr 219, http://biurose.sejm.gov.pl/teksty_pdf_03/r-219.pdf (dostęp: 19.09.2019).
Petrov Ch. (2019). *Big Data Statistics 2019*, https://techjury.net/stats-about/big-data-statistics/ (dostęp: 20.09.2019).
Popiołek M. (2018). *Czy można żyć bez Facebooka? Rola serwisów społecznościowych w sieciowym społeczeństwie informacyjnym*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.
Rosenberg E. (2018). *How YouTube Ad Revenue Works*, https://www.investopedia.com/articles/personal-finance/032615/how-youtube-ad-revenue-works.asp (dostęp: 25.09.2019).
Savov V. (2017). *YouTube's paid chats are an abomination from a capitalist dystopia I want no part of*, https://www.theverge.com/2017/5/18/15656908/youtube-super-chat-paid-abomination (dostęp: 27.09.2019).
Solon O. (2018). *YouTube's 'alternative influence network' breeds rightwing radicalisation, report finds*, https://www.theguardian.com/media/2018/sep/18/report-youtubes-alternative-influence-network-breeds-rightwing-radicalisation (dostęp: 27.09.2019).
Solsman J.E. (2017). *YouTube Super Chat lets you pay to spark livestream antics*, https://www.cnet.com/news/youtube-super-chat-superchat-api-water-balloons/ (dostęp: 24.09.2019).
Tyan I. (2019). *History teacher that teaches Nazi history has his YouTube account banned for "hate speech"*, https://reclaimthenet.org/history-teacher-that-teaches-nazi-history-has-his-youtube-account-banned-for-hate-speech/ (dostęp: 25.09.2019).
YouTube (b.d., a). *Wytyczne dla społeczności*, https://www.youtube.com/intl/pl/about/policies/#community-guidelines (dostęp: 21.09.2019).
YouTube (b.d., b). *Zasady dotyczące szerzenia nienawiści*, https://support.google.com/youtube/answer/2801939?hl=pl (dostęp: 21.09.2019).
YouTube (b.d., c). *Community Guidelines strike basics*, https://support.google.com/youtube/answer/2802032?hl=en (dostęp: 20.09.2019).
YouTube (b.d., d). *Jak zarządzać czatem na żywo*, https://support.google.com/youtube/answer/2524549?hl=pl (dostęp: 24.09.2019).
YouTube (b.d., e). *Super Chat eligibility, availability, & policies*, https://support.google.com/youtube/answer/9277801?hl=en (dostęp: 27.09.2019).
YouTube Official Blog (2011). *YouTube is going LIVE*, https://youtube.googleblog.com/2011/04/youtube-is-going-live.html (dostęp: 24.09.2019).
YouTube Official Blog (2019). *The Four Rs of Responsibility, Part 1: Removing harmful content*, https://youtube.googleblog.com/2019/09/the-four-rs-of-responsibility-remove.html (dostęp: 20.09.2019).